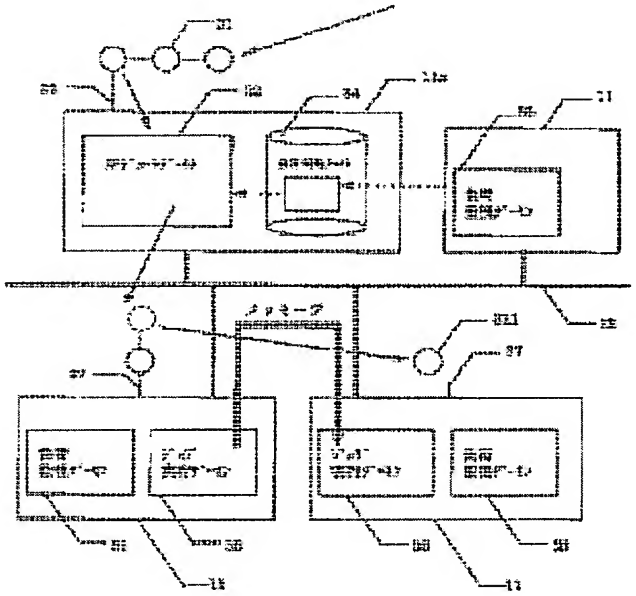
Ref. 1)

LOAD DISTRIBUTION CONTROL SYSTEM AND ITS DEVICE

Publication number: JP11312149 (A)
Publication date: 1999-11-09
Inventor(s): HIRATSUKA MASASHI
Applicant(s): HITACHI LTD
Classification:
- international: ***G06F15/16; G06F9/50; G06F15/177;*** **G06F15/16; G06F9/46;** (IPC1-7): G06F15/16
- European:
Application number: JP19980118122 19980428
Priority number(s): JP19980118122 19980428

Abstract of **JP 11312149 (A)**

PROBLEM TO BE SOLVED: To keep the balance of loads among computers by adding load information which indicates the load state of the self computer to a transmission message at the time of an inter-node communication processings among the optional computers and exchanging the load information between the computers. SOLUTION: Load information is transmitted from a load monitor demon 35 on the computer 11 to the computer 11a by a certain opportunity and is stored in plural pieces of load information file 34. Then, a scheduler demon 33 on the computer 11a determines by which one of the computers 11 a job or a process 31, which reaches a reception queue 32 is executed based on the pieces of load information, queues the job or the process 31 in the execution waiting queue 37 of the applying computer 11, successively takes-out it by a job execution demon and exchanges the pieces of load information between the respective computers 11. Thus, the balance of the loads among the respective computers 11 is kept.



(19)日本国特許庁（ＪＰ）　(12)公開特許公報（Ａ）　(11)特許出願公開番号

特開平11－312149

(43)公開日　平成11年(1999)11月９日

(51)Int.Cl.[6]	識別記号	ＦＩ	
Ｇ０６Ｆ　15/16	３８０	Ｇ０６Ｆ　15/16	３８０Ｚ

審査請求　未請求　請求項の数２　ＯＬ　（全　５　頁）

(21)出願番号　特願平10－118122

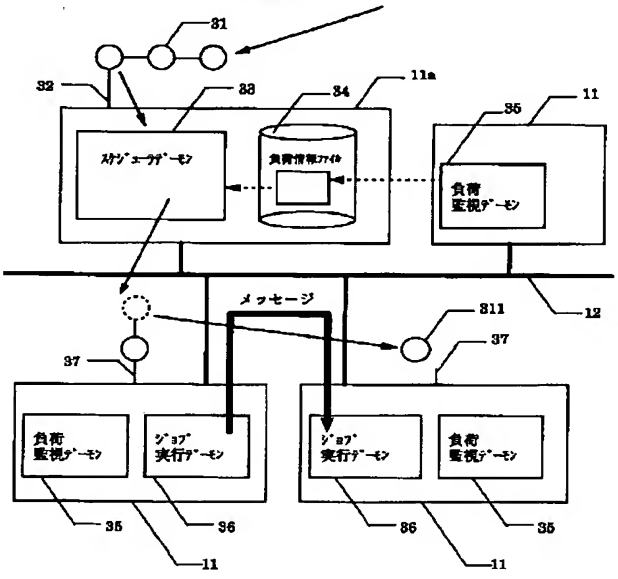
(22)出願日　平成10年(1998)４月28日

(71)出願人　000005108
株式会社日立製作所
東京都千代田区神田駿河台四丁目６番地
(72)発明者　平塚　正史
神奈川県横浜市戸塚区戸塚町5030番地　株式会社日立製作所ソフトウェア開発本部内
(74)代理人　弁理士　小川　勝男

(54)【発明の名称】　負荷分散制御方式及び装置

(57)【要約】
【課題】計算機の負荷情報に基づき、スケジューラ計算機がジョブまたはプロセスを計算機に割り当てる場合、時間が経過した負荷情報等が元で誤判断する場合があったが、計算機の正確な負荷情報を元に再割り当てが難しく効果的に計算機間の負荷の不均衡を是正することができなかった。
【解決手段】任意の計算機間でノード間通信処理を行う際、送信メッセージに自計算機の負荷状態を示す負荷情報を付加するようにすることで、計算機間で負荷情報の交換を図り、負荷の高い計算機に割り当てられた実行待ちジョブまたはプロセスを負荷の低い計算機に移動することによって、効果的に負荷の不均衡を是正する。

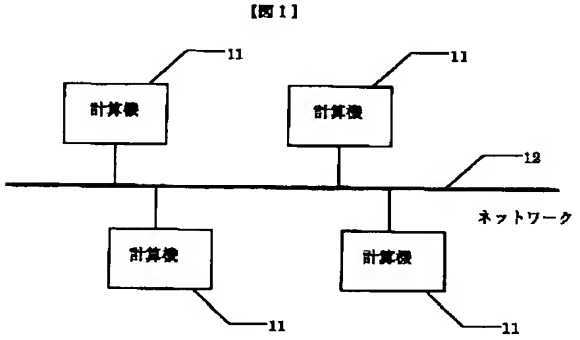
【図３】



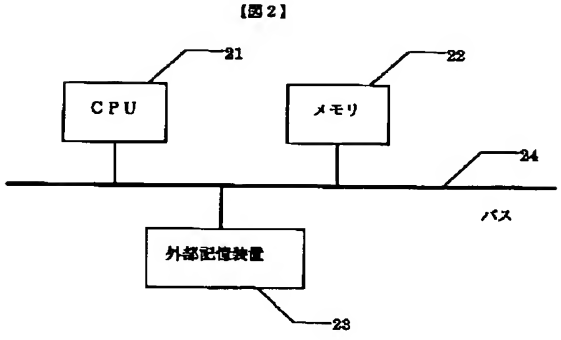
【特許請求の範囲】
【請求項１】ネットワークを介し複数の計算機が接続された並列計算機システムにおいて、ユーザから受け付けたプロセス及びジョブをどの計算機に実行させるべきかを判断するスケジューラデーモンを具備するスケジューラ計算機と、実行依頼されたプロセス及びジョブを単に実行する計算機（以下単に計算機といえばこちらを指す）から構成されるシステムにおいて、各計算機間でメッセージパッシング等のノード間通信（計算機間の通信処理）を行う際に、送信メッセージに自計算機の負荷情報を付加することにより計算機間で負荷情報を交換し、それにより計算機間の負荷の均衡を保つことを特長とする負荷分散制御方式。
【請求項２】複数の計算機がネットワークで接続されており、複数の計算機間でノード間通信を行う並列システムにおいて、各計算機に自計算機からの送信メッセージに自計算機の負荷情報を付加する装置と、受信メッセージに付加されている他計算機の負荷情報を認知するための装置と、それら負荷情報から負荷の高い計算機に割り当てられた実行待ちジョブまたはプロセスを負荷の低い計算機に移動する装置を具備することによって、効果的に負荷の不均衡を是正することを特長とする負荷分散制御装置。
【発明の詳細な説明】
【０００１】
【発明の属する技術分野】本発明は、並列計算機システムにおけるプロセス及びジョブの実行制御に係わり、計算機の負荷分散制御方式に関する。
【０００２】
【従来の技術】従来の方法の一例として、特開平０９－１６０８８４のように共有メモリを介して各計算機が他計算機と負荷情報を交換しあい、計算機間の負荷の均衡を図る方法があげられる。
【０００３】また上記に依らず、計算機の負荷情報を元に、ジョブまたはプロセスのスケジューリングをスケジューラ計算機が行う場合、各計算機からスケジューラ計算機に向けてのみ負荷情報の送信が行われていた。
【０００４】
【発明が解決しようとする課題】共有メモリを介して計算機の負荷情報を交換する方法は、共有メモリを有するシステムのみに適用可能であり、それ故適用範囲が限定されていた。
【０００５】また負荷情報を元に、スケジューラ計算機がジョブまたはプロセスをスケジュールする場合、実行以前に当該ジョブまたはプロセス実行時の計算機に与える負荷を正確に判断することが難しいため、結果的に誤った判断を下してしまう場合があり、それにより計算機間で負荷の不均衡が発生していた。
【０００６】なおこの負荷の不均衡が発生した場合、それを是正するために、負荷の不均衡が発生している計算機からスケジューラ計算機に負荷情報を送信する方法は、自計算機に更なる負荷を与えるだけでなく、スケジューラ計算機へのネットワークの局所的な負荷やオーバヘッドの増大を招き、これらの処理に時間を費やすことで送付された負荷情報の正確さが徐々に失われるため、スケジューラ計算機を介した方法では正確な負荷の不均衡是正の実現が困難であった。
【０００７】
【課題を解決するための手段】計算機間で負荷の不均衡が発生した場合、それを是正できるようにするため、任意の計算機間でノード間通信処理を行う際、送信メッセージに自計算機の負荷状態を示す負荷情報を付加するようにする。
【０００８】送信メッセージ及び負荷情報を受け取った計算機は、自計算機の負荷と、メッセージ送信側の負荷とを比較し、必要に応じて、負荷（未実行ジョブまたはプロセス）の移動を行うようにする。
【０００９】即ち、メッセージ送信側の計算機の負荷が高く、メッセージ受信側の計算機の負荷が低い場合には、送信メッセージに負荷が高い旨の負荷情報が付加され、メッセージ送信側計算機の未実行ジョブまたはプロセスを、メッセージ受信側計算機に移動するようにする。
【００１０】逆にメッセージ送信側の計算機の負荷が低く、メッセージ受信側の計算機の負荷が高い場合には、送信メッセージに負荷が低い旨の負荷情報が付加され、メッセージ受信側計算機の未実行ジョブまたはプロセスを、メッセージ送信側計算機に移動するようにする。
【００１１】なお負荷の高低は、両計算機間の負荷の相対的な比較に基づくが、負荷の差異が微少である場合には、両計算機がほぼ同一の負荷を有していると見なせるため負荷の移動には及ばない。よってこのように両計算機間の負荷の差異が微少と見なせるか、見なせないかの基準となる、差異許容値ｋを導入し、それによりメッセージ受信側計算機は、自計算機の方が負荷は高いか、低いか、または同一かを判断する。
【００１２】また未実行ジョブまたはプロセスを移動する際には、移動の旨及び、両計算機の最新の負荷情報をスケジューラ計算機に伝え、スケジューラ計算機は、当該計算機の負荷情報を入手することにより、それを元にスケジューリングを行うようにする。
【００１３】
【発明の実施の形態】以下、本発明の一実施例について図面により詳細に説明する。
【００１４】図１は本発明を適用した並列計算機システムのシステム構成を示すブロック図である。１１は計算機を示し、複数台の計算機１１がネットワーク１２にそれぞれ接続し、システムを構成する。図１では計算機は４台を示しているが、計算機１１の台数は任意である。
【００１５】図２は計算機１１の詳細な構成を示すブロ

は発生しない。
【００３５】(b)計算機Ｓは移動する未実行ジョブまたはプロセス３１１を自計算機の実行待ちキュー３７から切り離し、スケジューラデーモン３３から受け取ったジョブまたはプロセスの実行依頼内容を計算機Ｒに送付し、計算機Ｒは受け取った実行依頼内容を元に、自計算機の実行待ちキュー３７に未実行ジョブまたはプロセス３１１をキューイングする。
【００３６】(d)計算機Ｓ及び計算機Ｒは、共に自計算機の負荷が変化したことを伝えるため、自計算機の負荷をスケジューラデーモン３３に報告する。
【００３７】(iii)ケースＣの場合負荷の移動は発生しない。（図７、図８のケースＣ参照）
【００３８】
【発明の効果】共有メモリを有しない計算機システムにおいても各計算機の正確な負荷情報によりスクジューラ計算機が負荷の低い計算機へジョブヌはプロセスをスクジュールするためシステム全体のスループットを向上させることができる
【図面の簡単な説明】
【図１】本発明の一実施例を示すシステム構成ブロック図
【図２】図１の構成例に示した計算機の構成図
【図３】本発明の一実施例を示すシステム構成図
【図４】メッセージ送信側計算機とメッセージ受信側計算機との情報のやりとりを示す図
【図５】メッセージ送信側計算機とメッセージ受信側計算機との情報のやりとりを示す図
【図６】メッセージ送信側計算機とメッセージ受信側計算機との情報のやりとりを示す図
【図７】メッセージ送信側計算機のフローチャート図
【図８】メッセージ受信側計算機のフローチャート図
【符号の説明】
１１：計算機
１２：ネットワーク
２１：ＣＰＵ
２２：メモリ
２３：外部記憶装置
２３１：負荷情報ファイル
２４：バス
３１：ユーザから実行依頼されたジョブまたはプロセス
３１１：ユーザから実行依頼されたジョブまたはプロセスで、スケジューラ計算機から計算機を割り当てられたもの
３２：ジョブまたはプロセス受付けキュー
３３：スケジューラデーモンプログラム
３４：負荷情報ファイル
３５：負荷監視デーモン
３６：ジョブ実行デーモン
３７：実行待ちプロセスまたはジョブキュー

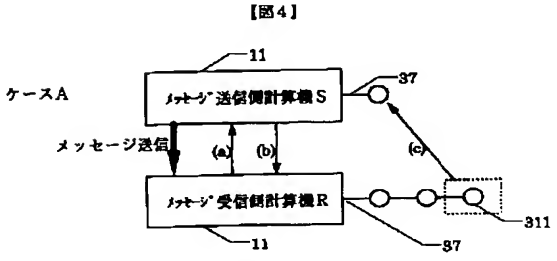
【図１】



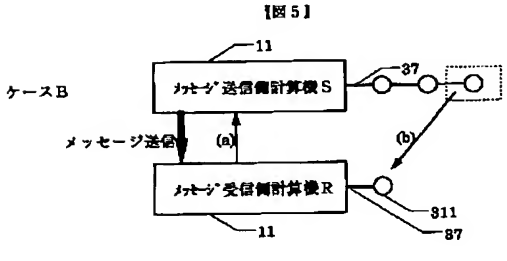
【図２】



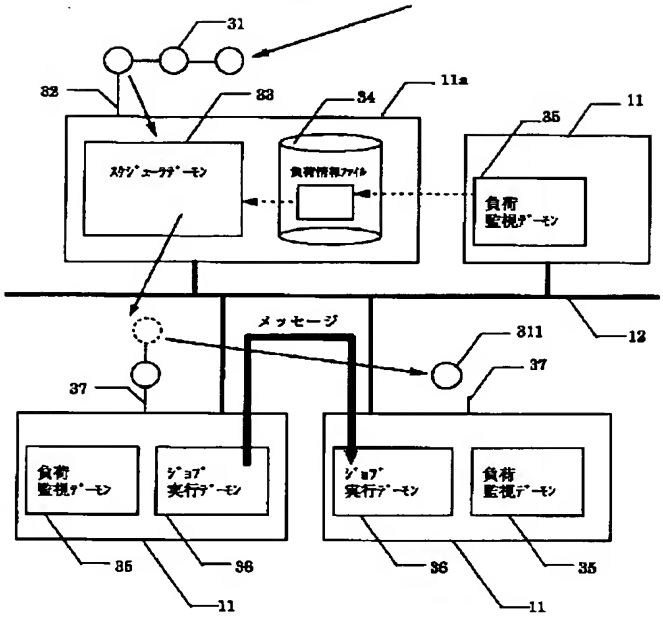
【図４】



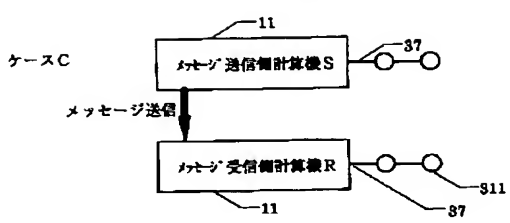
【図５】



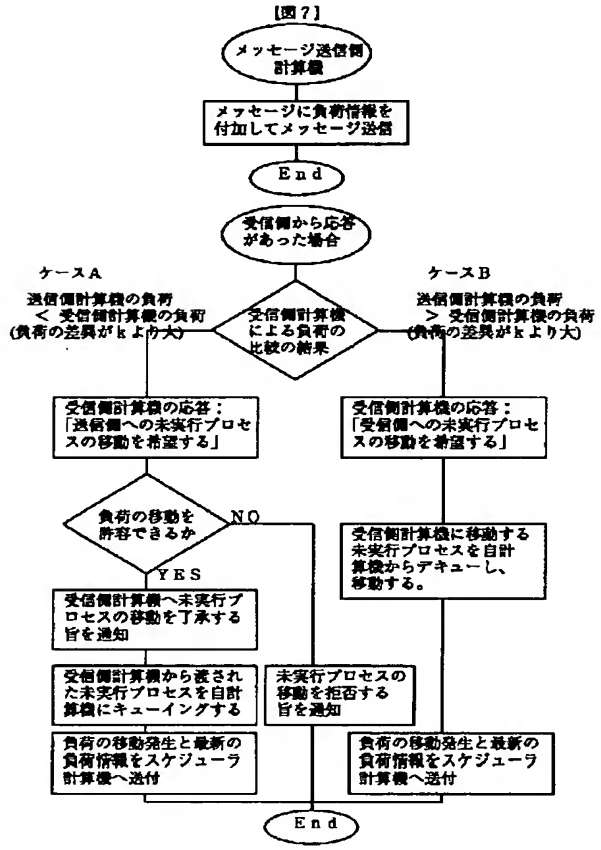
【図３】


31
32
33
34
11a
35
11
スケジューラデーモン
負荷情報ファイル
負荷
監視デーモン
メッセージ
311
12
37
37
負荷
監視デーモン
ジョブ
実行デーモン
ジョブ
実行デーモン
負荷
監視デーモン
35
36
11
36
35
11

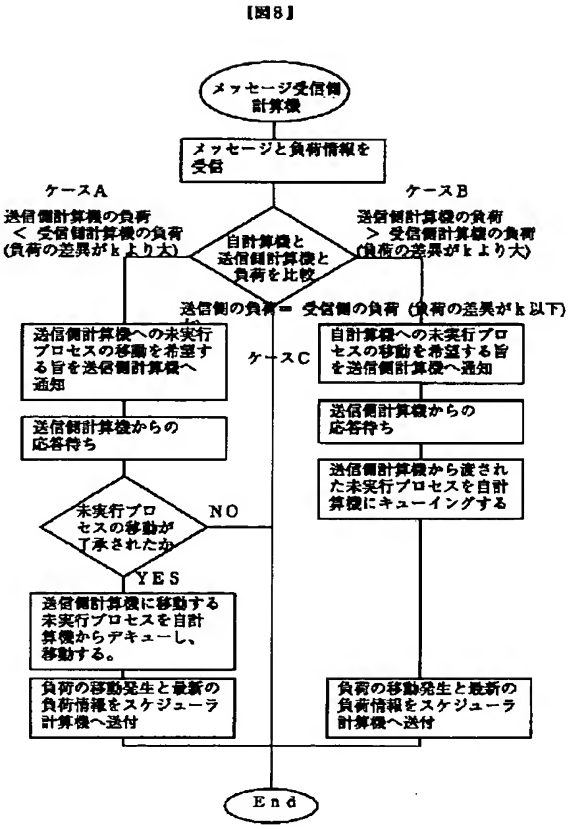
【図６】


11
ケースC
メッセージ送信側計算機S
37
メッセージ送信
メッセージ受信側計算機R
311
11
37

【図７】


メッセージ送信側
計算機
メッセージに負荷情報を
付加してメッセージ送信
End
受信側から応答
があった場合
ケースA
送信側計算機の負荷
＜ 受信側計算機の負荷
(負荷の差異がｋより大)
受信側計算機
による負荷の
比較の結果
ケースB
送信側計算機の負荷
＞ 受信側計算機の負荷
(負荷の差異がｋより大)
受信側計算機の応答：
「送信側への未実行プロセス
の移動を希望する」
受信側計算機の応答：
「受信側への未実行プロセス
の移動を希望する」
負荷の移動を
許容できるか
NO
YES
受信側計算機に移動する
未実行プロセスを自計
算機からデキューし、
移動する。
受信側計算機へ未実行プロ
セスの移動を了承する
旨を通知
受信側計算機から送ら
れた未実行プロセスを自計
算機にキューイングする
未実行プロセスの
移動を拒否する
旨を通知
負荷の移動発生と最新の
負荷情報をスケジューラ
計算機へ送付
負荷の移動発生と最新の
負荷情報をスケジューラ
計算機へ送付
End

【図８】


メッセージ受信側
計算機
メッセージと負荷情報を
受信
ケースA
送信側計算機の負荷
＜ 受信側計算機の負荷
(負荷の差異がｋより大)
自計算機と
送信側計算機と
負荷を比較
ケースB
送信側計算機の負荷
＞ 受信側計算機の負荷
(負荷の差異がｋより大)
送信側の負荷＝ 受信側の負荷 (負荷の差異がｋ以下)
ケースC
送信側計算機への未実行
プロセスの移動を希望する
旨を送信側計算機へ
通知
自計算機への未実行プロ
セスの移動を希望する旨
を送信側計算機へ通知
送信側計算機からの
応答待ち
送信側計算機から送ら
れた未実行プロセスを自計
算機にキューイングする
未実行プロ
セスの移動が
了承されたか
NO
YES
送信側計算機に移動する
未実行プロセスを自計
算機からデキューし、
移動する。
負荷の移動発生と最新の
負荷情報をスケジューラ
計算機へ送付
負荷の移動発生と最新の
負荷情報をスケジューラ
計算機へ送付
End